PhosphoGAN: Enhancing the Prediction Process of General and Kinase-Specific Phosphorylation Sites

Authors: *Mark Lennox, Prof. Adele Marshall, and Prof. Neil Robertson*
Affiliations: *Queen's University Belfast*

1. Abstract

Phosphorylation site prediction has now become a cornerstone within **protein function studies** and **experimental design**. Finding the exact location of these binding sites (aka **motifs**) is essential in a variety of domains such as **drug design and development**. To address this issue, several computational techniques have been developed in recent years that attempt to characterise and predict these sites via various forms of **feature extraction from raw protein sequence data**. When handling raw protein sequences, two central questions typically arise:

1. Firstly, can this particular protein sequence be **phosphorylated**?
2. Secondly, if this sequence can indeed be phosphorylated what **specific kinase** is the cause of this phosphorylation?

Figure 1:

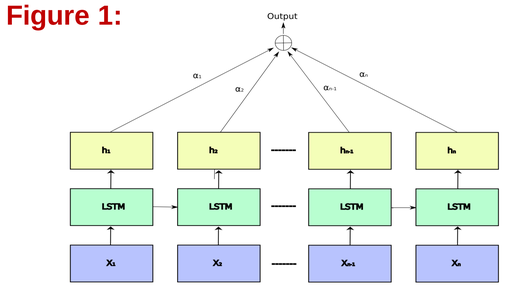
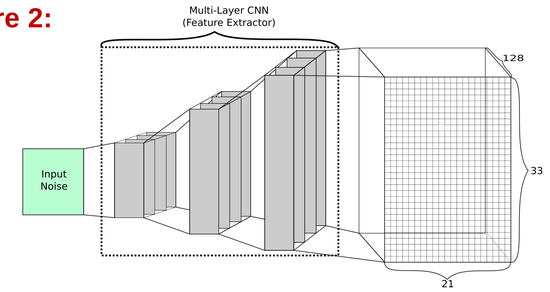


Figure 2:

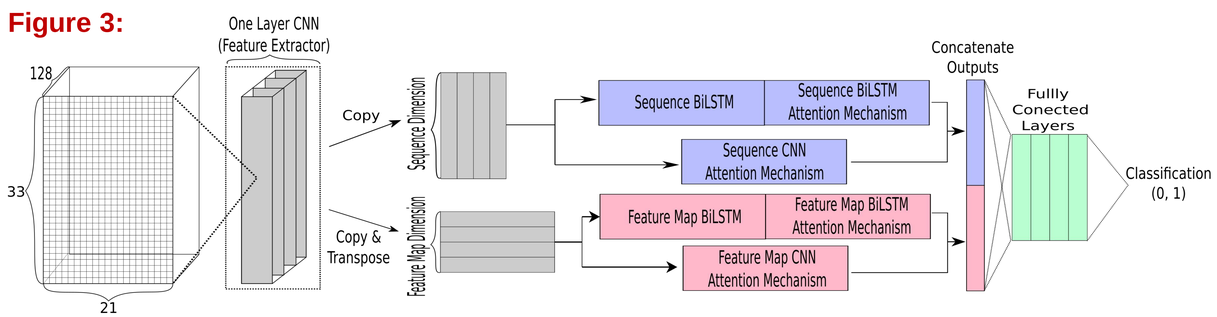


We present **PhosphoGAN**, a **semi-supervised generative adversarial approach** to training a deep neural network that is capable of **out-performing current state of the art models** such as **MusiteDeep** in predicting general and kinase-specific phosphorylation sites. **PhosphoGAN** produces a classifier via a semi-supervised approach using a generative adversarial neural network (GAN) to aid the training process as opposed to using such methods as transfer learning.

2. Methodology

The **discriminator** [3] used in PhosphoGAN begins with a set of convolutional layers (i.e. CNN). The output of the CNN is the copied and one version is transposed. Both copies are then passed through a dedicated **attention mechanism**. Accompanying each standalone attention mechanism, we have also introduced a pair of individual **bidirectional long short-term memory** (BiLSTM) models with their own independent attention mechanisms.

Figure 3:



The two-dimensional BiLSTM attention mechanism [1] (Att-BiLSTM) will analyse the feature maps produced by the CNN in both a **sequence and feature map dimension**. The outputs of the both Att-BiLSTM models and the stand-alone attention mechanisms are then concatenated together to form a set of features produced by the discriminator. These features are then passed through a series of full connected layers to attain a final classification.

3. Summary of Key Findings

The phosphorylation data used in the experiment is for Homo sapiens and was gathered from **UniProt/Swiss-Prot**. It consisted of the phosphorylation sites on **serine** (S), **threonine** (T) and **tyrosine** (Y), which provided a source for the positive data for the experiment, while the negative data was taking the same amino acid excluding annotated phosphorylation sites from the proteins.

General Classification Results:

Figure 4:

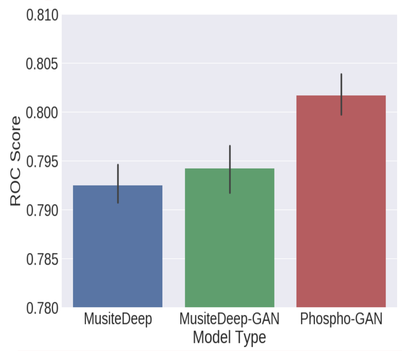


Figure 5:

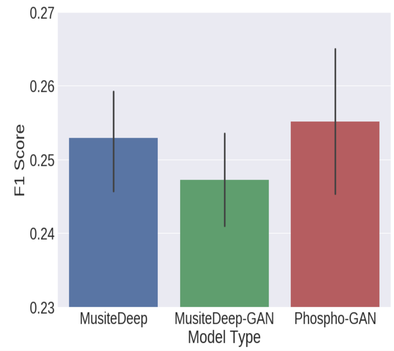
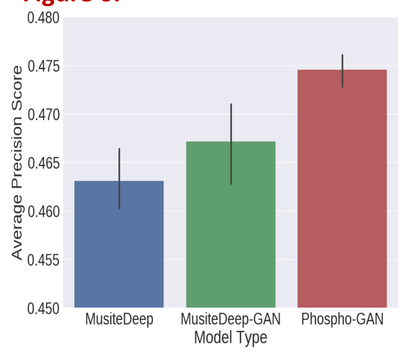


Figure 6:



To evaluate the performance of both models, a **five cross-fold validation** was used, and the area under the **receiver characteristic curve**, **average precision and F1 scores** were then calculated for each fold. Whereby **PhosphoGAN** outperformed MusiteDeep, and MusiteDeep-GAN in both general phosphorylation site and kinase-specific prediction.

Kinase-Specific Classification Results:

Figure 7:

Model	ROC	F1 Score	Average Precision
CDK-MusiteDeep	0.84	0.1564	0.469
CDK-MusiteDeep-GAN	0.8824	0.194	0.5025
CDK-PhosphoGAN	**0.9016**	**0.2488**	**0.5183**
CK2-MusiteDeep	0.8156	0.2211	0.4508
CK2-MusiteDeep-GAN	0.8152	0.2279	0.4513
CK2-PhosphoGAN	**0.8344**	**0.2373**	**0.4711**
MAPK-MusiteDeep	0.7671	0.1506	0.4793
MAPK-MusiteDeep-GAN	0.8694	0.23	0.4958
MAPK-PhosphoGAN	**0.8793**	**0.2384**	**0.5058**
PKC-MusiteDeep	0.7918	0.132	0.4294
PKC-MusiteDeep-GAN	0.7994	0.1324	0.4368
PKC-PhosphoGAN	**0.8045**	**0.1419**	**0.4373**

4. Conclusion

By applying a new semi-supervised training approach along with a new model architecture for the classifier, we obtain results that **outperform the current state of the art MusiteDeep model**. These results demonstrate how deep learning can be applied with significant effect to a problem where the training data is insufficient and unbalanced.